基于科学数据生命周期管理阶段的科学数据质量评价体系构建研究

■ 江洪[1] 王春晓[1,2]
[1] 中国科学院武汉文献情报中心 武汉 430071 [2] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190

摘 要：[目的/意义]选取国内外 15 家科学数据中心的科学数据质量评价指标，旨在筛选能够客观反映科学数据质量的共性指标，构建具有普适性的科学数据质量评价指标体系。[方法/过程]采用文案调查法、网络调查法和内容分析法，对 15 家科学数据中心的科学数据评价指标进行梳理和分析，了解现有的科学数据机构的数据评价指标。[结果/结论]基于科学数据生命周期管理的各个阶段构建一套由数据管理计划、数据收集管理、数据分析与加工管理、数据保存管理和数据共享利用管理 5 个维度组成的科学数据质量评价指标模型，为我国和地方科学数据中心建立面向决策的科学数据中心评价系统提供参考。

关键词：科学数据 数据生命周期 评价体系 指标
分类号：G203
DOI：10.13266/j.issn.0252-3116.2020.10.003

chinaXiv:202304.00241v1

当今社会，随着科学研究的不断发展，科学数据的数量变得越来越庞大，结构也越来越复杂。科学数据具有巨大的科研价值，对科学数据的研究成为科学研究的重中之重。科学数据的评价是科学数据管理和服务机构都要重视的关键环节。国外科学数据机构在评价方面已经做了大量工作，部分经验值得国内科学数据中心借鉴。例如美国国家海洋和大气管理局的信息质量指南中提出了可用性、客观性、完整性、影响力、透明度、再生性等指标和这些指标的使用范围[1]；荷兰数据存档和网络服务(Data Archiving and Networked Services，DANS)对其在线存储系统上的数据集进行评估，评估指标有可发现性(findability)、可达性(accessibility)、互操作性(interoperability)、可重用性(reusability)[2]，这也是科学数据管理中的 FAIR 准则[3]。我国 2008 年开始要求国家项目产生的科学数据进行汇交，相继出台了各种项目的科学数据汇交办法[4]。2018 年国务院出台的《科学数据管理办法》更体现出国家对科学数据这一战略资源的重视。制定合理的科学数据质量评价指标体系能够促进我国国家和地方科学数据中心建设，有利于我国逐步建设知名的科学数据评价研究中心。但目前国内对科学数据的评价方面研究较少，科学数据的评价缺乏统一的标准。本文通过调研国内外科学数据机构的科学数据质量评价指标，试从科学数据生命周期管理视角建立科学数据质量评价模型，以期能为国内科学数据评价的相关研究和工作提供参考。

1 国内外科学数据质量评价研究现状

1.1 国外研究现状

从广义的数据质量来看，国外学者对数据质量评价关注得较多，构建了众多的数据质量评价模型：如 B. Stvilia 等从内在信息质量、情境信息质量及信誉信息质量 3 个维度出发构建信息质量评价模型[5]；C. Batini 等提出基于方法论的数据质量维度，包括完整性、准确性、及时性、一致性、可访问性、可信性、可用性、可解释性和适当的数据量等[6]；A. Zaveri 等构建了 18 个不同的数据质量维度来评价关联数据，并将这些数据质量维度分为 4 组：可访问性维度、情境维度、本征维度、表征维度[7]。有部分学者构建了针对某一学科领域的科学数据质量评价模型：如 M. G. Kahn 等构建了针对电子健康记录临床研究数据的质量评价模型，该模型主要指标有准确性、可信性、客观性、及时性和数

作者简介：江洪(ORCID:0000-0003-3806-1856)，副主任，研究员，硕士生导师，E-mail:jianghong@mail.whlib.ac.cn；王春晓(ORCID:0000-0002-2131-5111)，硕士研究生。

收稿日期：2019-10-29 修回日期：2020-02-09 本文起止页码：19-27 本文责任编辑：易飞

ChinaXiv合作期刊

据量的合理性[8];H. Chen 等构建了 3 个维度的数据质量评价模型来评估公共卫生领域相关数据,这 3 个维度分别是数据本身、数据使用和数据收集过程[9];H. Huang 等在前人数据质量评价标准的基础上提出了基因组注释环境中适用的数据质量标准[10]。调研发现,不同学者提出的数据质量维度虽然繁多,但有交叉重复的内容。当评价某一种具体科学数据时,在不同的使用情境中,数据质量的维度具有不同的优先顺序。现阶段学者们多是关注某一学科领域的数据或技术平台的评价研究。

1.2 国内研究现状

国内已有不少对数据的评价研究,其中涉及政府开放数据的评价居多,例如邵艳红根据已有的评价指标和数据质量标准构建政府开放数据质量评价标准[11];李晓彤等在北京、广州和哈尔滨三市超过 1 900 个数据集的质量问题调查的基础上,提炼出 7 个质量维度和可度量的评价指标,分别是完整性、时效性、一致性、准确性、唯一性、可理解性和开放性[12]。有不少学者关注科学数据平台建设的评价,如刘桂锋等分析了 6 个国际组织开放政府数据的评估项目,提取出适用于科学数据平台的指标,从平台建设基础、平台管理功能、平台数据及平台效果与影响 4 个方面构建科学数据平台评价指标体系,并在指标体系中结合了数据生命周期理论来构建二级指标[13];周宇等通过调研国内外数据监护平台,并采用专家调查法,最终确定了数据监护平台的评价指标体系,包含数据管理制度、服务功能、数据量、数据质量、平台界面、软件系统及利用率等维度[14]。除了关注科学数据平台的评价之外,有些研究者关注数据质量本身的评价,如余芳东从数据源条件、元数据、数据质量 3 个方面构建指标框架来评价政府统计数据中的非传统数据[15];余厚强等通过梳理替代计量数据生产流程,构建了替代计量数据质量评估体系[16]。国内对科学数据质量评价的研究内容较少。目前,国内科学数据机构出台的关于科学数据质量的评估体系几乎只涉及准确性、完整性和可用性等宽泛的指标。本文试图在考虑数据生命周期的基础上,通过调研不同学科领域的科学数据机构,构建适用于科学数据管理生命周期的不同阶段的质量评价模型。

2 研究方法

马费成和望俊成认为,生命周期方法适用的对象应该具备 3 个重要的属性——“连续性、不可逆转性和迭代性”,生命过程的不同阶段之间不仅具备连续性,而且具备时间上的不可逆转性,完成一次生命进程后,会进入下一轮生命进程,两轮之间的更迭也就是迭代或循环[17]。根据这一理论,丁宁等提出生命周期方法也可适用于科学数据中,科学数据生命周期与科研流程密切相关,科学数据生命周期管理的本质是依据科研工作流程管理数据[18]。不同的科研活动可能只包含科学数据生命周期中的部分阶段,例如一个主要关注数据处理和分析的科研项目可能会绕过数据产生、采集等阶段[19]。从科学数据生命周期管理的视角来分析数据评价指标,能够在指标体系中更明显地体现出依据科研流程进行科学数据管理的特征,有利于更有效地进行科学数据生命周期管理。张洋和肖燕珠通过对 10 种数据生命周期理论进行调研分析,总结出了科学数据生命周期管理的 5 个核心阶段,分别是制定数据管理计划、数据收集管理、数据分析与加工管理、数据保存管理、数据共享与利用管理[20]。本文以科学数据生命周期管理的 5 个阶段作为维度,从这 5 个阶段来分析科学数据的具体评价指标。

本文采用文案调查法、网络调查法和内容分析法,在进行广泛的网络调研的基础上,选取了 15 家有明确提出指标的数据机构(见表 1),主要分布于美国、欧洲以及中国。调研过程中笔者主要关注该数据机构在科学数据管理计划、数据收集管理、数据分析与加工管理、数据保存管理、数据共享与利用管理 5 个阶段的评价指标。这 15 个数据机构的科学数据内容涉及地理、生物、医药卫生、社会、经济以及其他自然科学领域,内容比较全面。有些数据机构的数据资源集中在一两个学科领域,有些数据机构则关注众多学科领域,其数据资源较为丰富。

3 科学数据质量评价指标分析

3.1 制定数据管理计划

《科学数据管理办法》强调法人单位和各级主管部门制定好科学数据管理计划,并履行科学数据管理的职责[21]。该阶段主要任务是计划好如何描述和存储数据,即有完整的元数据标准,例如定义数据类型、格式等;以及在整个数据生命周期过程中如何管理、访问和共享数据,如规定数据管理的职责分配,确保有相应的专业人员来执行数据管理计划。

从科学数据管理计划可以看出科研人员和研究组织在数据管理方面的意识和能力,对调研结果进行整理分析后发现,这一阶段的指标内容主要与数据管理

chinaXiv:202304.00241v1

ChinaXiv合作期刊

表1 15家科学数据机构基本情况

机构编号	名称	责任者	数据类型	学科领域	数据开放网址
1	美国国家海洋和大气管理(National Oceanic and Atmospheric Administration,NOAA)	美国国家环境信息中心	科技数据	地球物理、地质、气象及环境科学等	https://www.cio.noaa.gov/services_programs/info_quality.html
2	美国地球资源观测与科技中心(Earth Resources Observation and Science,EROS)	美国地质调查局	科技数据	地质学	https://www.usgs.gov/centers/eros/data-tools
3	橡树岭国家实验室分布式活动档案中心(The Oak Ridge National Laboratory Distributed Active Archive Center,ORNL DAAC)	美国宇航局	科技数据	环境、生态	https://daac.ornl.gov/
4	德克萨斯数据仓储(Texas Data Repository,TDR)	德克萨斯数字图书馆	科技数据、社会、经济数据等	综合	https://data.tdl.org/
5	纽斯卡尔大学开放数据存储库(Newcastle University Open Data Repository,NCL Data)	纽斯卡尔大学	科技数据	综合	https://data.ncl.ac.uk/
6	社会科学数据档案(Social Science Data Archive,SSDA)	加州大学洛杉矶分校图书馆	科技数据	社会科学	https://dataverse.harvard.edu/dataverse/ssda_ucla
7	大气辐射测量(Atmospheric Radiation Measurement,ARM)	美国能源部	科技数据	综合	http://adc.arm.gov/discovery/# v/home/s/
8	深蓝数据(Deep Blue Data)	密歇根大学	科技数据	综合	https://deepblue.lib.umich.edu/data/?locale=en
9	通用蛋白质资源知识库(Universal Protein Resource Knowledgebase,UniProtKB)	欧洲生物信息研究所	科技数据	医药卫生	https://www.uniprot.org/
10	地球数据观测网(Data Observation Network For Earth,DataONE)	美国国家科学基金会	科技数据	综合	https://www.dataone.org/
11	国家基因库生命大数据平台(China National GeneBank DataBase,CNGBdb)	深圳国家基因库	科技数据	生物	https://db.cngb.org/datamart/animal/
12	国家基因组科学数据中心(National Genomics Data Center,NGDC)	中国科学院北京基因组研究所	科技数据	生物	https://bigd.big.ac.cn/databases
13	明尼苏达大学数据仓储(Data Repository for University of Minnesota,DRUM)	明尼苏达大学	科技数据	综合	https://conservancy.umn.edu/discover
14	英国数据档案(UK Data Archive,UKDA)	埃塞克斯大学	综合	综合	https://beta.ukdataservice.ac.uk/datacatalogue/studies/#!?Search=&Page=1&Rows=10&Sort=0&DateFrom=440&DateTo=2019
15	世界数据系统(The World Data System of the International Science Council,ICSU-WDS)	国际科学理事会	科技数据	地球物理	http://www.icsu-wds.org/services

chinaXiv:202304.00241v1

计划(Data Management Plan,DMP)的制定有关。其中包括以下4个方面:①DMP的完整性。ORNL DAAC强调提供的DMP要尽可能包含描述数据所需的内容,例如对特定类型数据要有精度和密度的合理说明[22];SSDA在计划阶段认为应当创建全面的数据文档来解释数据是如何被创建的[23];ARM规定DMP必须描述数据如何共享和保存,并包含个人隐私和机密信息方面的要求[24];EROS考虑了数据管理的过程完整性和预算开支等[25]。②数据管理职责。EROS在这阶段的指标说明比较典型,其认为应当有专业人员进行数据管理;管理职责范围应当明确;应当符合机构官方的要求;确保开发和维护元数据在内的数据文档;制定数据质量标准[26]。③DMP的价值性。SSDA和ARM强调DMP的价值性,前者认为DMP应当能够赢得资助者的信服和支持[27];后者认为DMP要具有助力科研的价值,并应当通过其机构的数据价值审核程序[24]。④DMP制定是便利的、易操作的。NCL Data强调要提供多种DMP格式以供参考;提供制定DMP的培训、指南或帮助;提供创建DMP的链接;创建DMP过程中提供联系方式以供咨询[28]。

在制定计划阶段对DMP本身的评价是一个处于演变中的新概念[29],一份好的DMP文件其内容应当对数据生命各个周期要注意的事项进行说明,并体现具体研究项目及资助机构的要求。所调研机构对这一部分内容提及有限。

3.2 数据收集管理

我国各级科技部门和科研人员逐渐认识到科学数据的重要价值,科学数据是新一轮科技创新的重要基础。建设科学数据中心离不开对科学数据的收集。《办法》强调由各法人单位承担其相关领域科学数据

ChinaXiv合作期刊

的整合汇交工作,各单位应当有科学数据的质量控制体系来保证数据表达的准确性和数据可用性[21]。经调研后分析整理该阶段对科学数据的评价指标主要有以下5个方面:

(1)数据收集的格式要求。Deep Blue Data 和 ORNL DAAC 建议提交的数据采用非专有格式、开放格式[22,30]; UniProtKB 建议使用符合 UniProtKB 要求的数据格式[31];DRUM 认为提交的数据应当符合其给定的适合访问的文件格式,并且不同的数据类型有不同的格式规定[32];TDR 对表格数据文件有格式要求,要求提交 SPSS(POR 和 SAV 格式)、STATA、R data、CSV 等格式[33];NGDC 也强调数据提交要采用规定的标准格式[34]。

(2)数据审核。CNGBdb 标明所提交的数据需要通过 MD5 校验数据传输的完整性,而且需要通过元数据信息和伦理批件等审核[35-36];ORNL DAAC 强调要审核数据的优先领域、科学影响和社区需求来确定数据的优先级[37]。

(3)对数据内容的要求。这部分指标主要关注所收集数据的相关性、完整性和准确性。相关性指标包含所选数据是否被判断为主题相关、是否有相关性判断标准;完整性指标包含数据有完整的元数据描述以及内容的完整性,包括 DRUM 将数据按照时间和相关性进行排序、Deep Blue Data 强调元数据的完整性[30]、TDR 要求元数据的描述符合标准、完整不漏[38];UKDA 强调检查测量数据的准确度,使用多次测量、观察或取样以及专家核对等方法来确保数据的准确性,还提到数据和元数据的数字化程度[39]。

(4)数据表达。ORNL DAAC 强调数据描述清晰易于理解[37];UKDA 建议在收集过程中尽量使用受控词汇,减少手工输入[39]。

(5)数据重复使用。这部分指标主要内容有数据利用的可重用性和可复制性。Deep Blue Data 建议数据包含描述性元数据,应当能被他人重复使用[30];DataONE 建议研究成果可以被他人复制[40];DRUM 表示所有数据都要接受审查以确保能重新使用,没有重用功能的数据可能不会被存储库接受[41]。

3.3 数据分析与加工管理

数据的分析与加工处理是指利用数据处理软硬件资源,针对用户的需求,对有关数据进行加工或分析处理,并将得到的数据加工产品和分析处理结果以合适的方式提供给用户的服务。科学数据的分析与加工目的是挖掘和提升科学数据的产品价值,使科学数据具有可发现性(或可查找性),可访问性、增值性、互操作性等[20]。

所选取的15家数据机构中,有6家提到了科学数据分析与加工方面的指标,其具体指标内容见表2。经过整理后发现,该阶段的指标主要可以归纳为4个方面:①数据创建与描述。其内容有创建元数据的方法和标准、开发数据字典、文件名称具有有效性等。②数据处理。其内容有数据的加工深度、加工效率;对数据的分类;数据更新是否及时;是否能够可视化处理等。③数据可发现性。其内容有处理数据的代码可以共享;数据是易于检索的;文件名是有效的等。④数据可利用性。其内容有处理数据的代码可供他人使用;数据集具有增值性等。

3.4 数据保存管理

科学数据的长期保存要求存储库具有很高的安全性,不同的科学数据集在安全性指标方面有不同的内容。数据保存对存储库系统有技术上的要求,对存储内容本身也有要求,包括保存格式和数据内容的机密性、完整性和可用性等。所调研的数据机构在这方面的指标可以划分为以下5个方面:①数据保存安全性。Deep Blue Data 提到存储设施应当具有适当灾难恢复功能,提供比特级保护;强调数据转移过程中的完整性和安全性[30];EROS 提到系统的安全性要求,并提出应当明确谁来负责 IT 安全和隐私,另外强调安全协议的重要性不能忽视[46];TDR 提出科学数据安全性必须考虑到数据的备份,定期检查,提供资源服务密钥[47]。②数据保密性。NCL Data 提到在存储数据时应当采用文件加密技术[48]。③数据保存的易操作性。易操作性强调用户与系统的互动,体现在用户遇到困难时是否能获取帮助。例如 NCL Data 提到遇到存储问题时可提交解决申请。④数据存储内容指标。经调研发现该部分指标主要关注数据内容是否仍然可以访问;存储内容及存储系统是否及时更新;数据是否仍然具备利用价值;数据内容是否能长期保存;数据存储量大小的规定;提供不同级别的数据保存服务等。⑤数据保存格式。对数据保存格式的要求,格式是否具有可移植性以及多样性。例如 EROS 认为数据存储格式应当具有可移植性,多年以后仍然可以使用[49]。

3.5 数据共享与利用管理

国外积极推动科学数据共享的主要动力有:①推动科学研究;②避免重复研究造成资源浪费;③ 有效长期保存科学数据;④促进科学研究的合作,提高科研成果的引用率和影响力[50]。我国尚未形成有效的数

chinaXiv:202304.00021v1

ChinaXiv合作期刊

表 2　科学数据机构在数据分析与加工管理阶段的指标

数据机构	评价指标	指标说明
NOAA	加工程度	对数据被加工的深度进行评价
	数据分类	有无数据分类,分为哪些类
	及时性	更新的及时性
	创建元数据	是否给出创建元数据的方法[42]
EROS	数据加工效率	使用脚本语言自动化处理和简化文档;是否有其他提高处理效率的措施
	可读性	代码可读性、文档易懂;应当支持开放源码软件开发
	可重用性	允许别人重新运行你的分析;代码建议有版本控制
	可达性(accessibility)	处理的代码是否放到公共存储库;确保数据是可用的
	文档管理指标	应当维护数据处理和分析活动的文档;应当有帮助这一阶段流程再现的补充材料
	数据标准指标	应当创建并依据数据标准;数据文件的创建格式要求;数据组织应当有逻辑且易于发现和访问
	可发现性或易于检索性	应当使用与数据集相关的关键字;文件名称是否具有一致性[43]
ORNL DAAC	代码可读性	在代码中编写注释,利于其他人使用
	可复制性	处理的数据的代码是可以复制给别人使用的
	及时性	数据应当及时更新
	数据字典	开发明确定义参数、属性、变量的数据字典
	互操作性	遵循为数据互操作性建立的标准,如气候和预测(CF)元数据约定[22]
TDR	便利性	提供内置的数据可视化工具及其使用指南[44]
ARM	及时性	及时甚至实时将数据处理的结果反馈给相关人员;定期处理、整理和存档电子仪器现场数据
	描述完整性	对仪器(或系统)、VAP 技术、QME 技术或其他方法的完整描述
	有效性	文件名的有效性;算法高效
	准确性	对科学数据的一些概念进行明确的定义[45]
UKDA	可扩展性	研究人员可以通过添加额外的变量或参数来扩展可能的应用程序,从而为他们的数据集增加重要的价值[39]

据开放机制,各个政府部门、科研机构之间的数据共享仍然存在壁垒,形成了“数据孤岛”[51]。经调研发现,所调研的机构在该阶段涉及到的指标内容主要有数据发布、数据引用、数据开放程度、数据影响力、数据使用的合法性以及数据共享的隐私问题等 6 个方面:①数据发布。EROS 强调要有发布格式要求;发布的产品要包含规定的要求,即具有完整性要求;发布之前应当审查数据的准确性、一致性和完整性等[52]。②数据引用。这方面主要关注 DOI 问题,如是否有统一的数据引用标准或规范;是否为数据分配 DOI;应当符合数据引用格式;给出引用数据的指南或帮助等。③数据开放程度。ICSU-WDS 认为,为公共领域使用的数据、元数据、产品和信息应根据响应的法律法规充分实现公开共享[53];ARM 和 DRUM 皆支持免费开放。④数据影响力。NOAA 强调数据受众范围和传播的及时性;并评估其是否对重要的公共部门或企业的决策产生实质影响[42]。⑤数据使用合法性和数据共享的隐私问题。TDR 强调数据使用者不得侵犯他人权利;尊重他人隐私;遵守所有适用的当地、州、国家和国际法律及德克萨斯数字图书馆使用协议规定[54];ICSU-WDS 强调数据应当符合国际伦理行为研究标准;遵守国家或国际法律和政策;数据共享应当确保一定的隐私;适当情况应当标记敏感信息或受限制的信息[53]。

4　科学数据质量评价模型构建

笔者从科学数据管理的生命周期各阶段的视角,对调研结果进行分析归纳,从制定数据管理计划、数据收集管理、数据分析与加工管理、数据保存管理和数据共享与利用管理 5 个维度,结合系统性、科学性、简明性、通用性和可操作性等评价指标体系应有的要求,构建了科学数据质量评价模型(见表 3)。本文的指标是基于所调研的 15 家机构的做法进行抽象归纳的,未有超出这些机构做法的指标(相关机构已在表 3 中注明)。该指标体系共分为 3 个指标层次,所构建指标是总-分的逻辑关系,力求充分体现科学数据管理生命周期各阶段的特性。同时,笔者在调研过程中发现,可用性、完整性和客观性这 3 个指标,是贯穿科学数据管理生命周期各阶段的科学数据质量评价的共同指标,NOAA 对这 3 个指标的指标说明较为典型(见表 4)。在对科学数据质量进行评价时,不仅要考虑每个生命周期阶段的个性指标内容,更要结合可用性、完整性及客观性这 3 项内容。因此在构建评价模型的过程中也结合了这 3 个指标来制定每一个指标的具体评价内容。

chinaXiv:202304.00241v1

ChinaXiv合作期刊

表 3 科学数据质量评价指标体系

一级指标(机构编号)	二级指标(机构编号)	三级指标	评价指标描述
制定数据管理计划(2,3,5,6,7)	A1 DMP 文件的制定(3,5,6,7)	A11 DMP 的完整性	DMP 包含符合要求的所有所需信息的程度
		A12 DMP 的准确性	DMP 语句描述清晰,内容准确的程度
		A13 DMP 制定的易操作性	DMP 制定流程的便利程度(例如提供制定 DMP 的指南、培训或帮助以及工具链接等)
		A14 DMP 的价值性	DMP 获得资助者认可和信服的程度
		A15 DMP 的规范性	符合 DMP 制定标准和格式要求的程度
	A2 数据管理职责(2)	A21 数据管理的专业性	专业人员参与数据创建或管理的程度
		A22 管理职责明确性	要明确数据管理的工作内容和职责范围
数据收集管理(3,4,8,9,10,11,12,13,14)	B1 数据提交过程(3,4,8,9,10,13)	B11 数据格式规范性	提交格式符合系统要求的程度
		B12 数据格式多样性	为不同数据类型提供不同的数据格式
		B13 数据提交便利性	数据所有者进行数据提交的便利程度
		B14 可重用性	所提交的数据能被他人复制使用的程度
	B2 数据内容(4,8,11,12,13,14)	B21 数字化程度	所收集的数据的数字化水平
		B22 数据完整性	数据传输的完整程度和数据描述的完整程度
		B23 数据内容准确性	数据内容的准确、真实程度
		B24 相关性	所收集的数据内容与系统要求的主题的相关程度
	B3 数据表达(3,14)	B31 可理解性	数据的表达在多大程度上能使用户理解以及机器可读程度
		B32 数据表达的规范性	数据表达的标准化程度(例如是否使用受控词汇)
数据分析与加工管理(1,2,3,4,7,14)	C1 数据创建(1,2,3)	C11 元数据创建标准化	元数据创建过程有标准可循
		C12 数据文件有效性	所创建的数据文件的格式、名称的有效程度
		C13 数据字典	应当开发创建数据字典
	C2 数据加工与处理(1,2,3,4,7,14)	C21 数据加工深度	对数据的加工深度进行评价
		C22 数据加工效率	数据加工的速度和数据量的大小
		C23 互操作性	数据在多大程度上遵循数据互操作性标准
		C24 数据可视化	系统能够为数据提供可视化处理的程度
		C25 数据更新及时性	数据能够及时更新和维护的时间周期
		C26 可扩展性	能够在多大程度上添加额外的措施来扩展可能的应用程序,从而增加数据集的价值
数据保存管理(2,4,5,8)	D1 存储系统(2,4,5,8)	D11 安全性	系统能提供数据安全保存的程度
		D12 可迁移性	数据在受到安全威胁时可以迁移的程度
		D13 可恢复性	系统具备的灾难恢复程度
		D14 保密性	数据符合系统要求的保密程度
		D15 一致性	数据的属性在不同系统中相符合的程度
		D16 存储量	系统能提供多大的数据存储容量
		D17 格式规范性	符合系统要求的数据保存格式的程度
		D18 持久性	数据内容能够在多长时间范围内完整并可持续地保存
	D2 存储操作(4,5)	D21 数据备份	应当定期检查和安全备份
		D22 便利性	数据存储过程中提供帮助、指南或培训的程度
		D23 解决问题的效率	数据存储过程中解决问题的有效程度
数据共享与利用管理(1,2,4,7,13,15)	E1 数据共享(4,7,13,15)	E11 数据开放程度	数据在多大程度上支持用户开放获取
		E12 合法性	数据的共享应当尊重他人权利,并遵守相关法律法规
		E13 免费性	数据可以免费获取的程度
		E14 隐私性	符合数据共享对隐私保护的要求的程度
		E15 数据发布规范性	数据产品的发布符合规范要求的程度
	E2 数据利用(1,2,15)	E21 可访问性	用户能获得的数据访问权限的程度以及在特定环境中数据可以访问的程度
		E22 可引用性	数据能被用户规范性引用的程度,应当为数据分配 DOI
		E23 数据利用率	数据被访问、下载和使用的情况
		E24 数据影响力	数据传播的广泛程度和对决策产生的实质影响

chinaXiv:202304.00241v1

ChinaXiv合作期刊

表 4　贯穿科学数据管理生命周期各阶段的科学数据质量评价的共同指标

共同指标	指标说明
可用性	以用户为中心、容易访问、容易阅读和理解、高透明度、提供数据背景资料、适应各种操作系统[1]
客观性	数据准确性(信息的不精确性或误差在可接受的范围内,且符合通常接受的科学、财务和统计标准)、来源可靠、数据描述清晰、数据可追朔[1]
完整性	数据的完整性不被不适当的访问所修改、破坏;符合内部安全标准[1]

5　结语

构建科学数据质量评价指标体系是科学数据评价和管理的重要工作,本文构建的指标体系考虑到数据生命周期各阶段的特征和目标,以期能为科学数据机构平台建立科学数据评价体系提供参考和补充。因为该指标体系涉及到科学数据生命周期管理的各个阶段,在实际操作中,科学数据机构可以根据所辖数据的类型、特征和数据管理要求等来具体借鉴相应的指标构建适用于本机构的评价指标。本文构建的指标体系尚存在实验数据不充分、分析不够系统等问题,下一阶段的研究目标则是充分论证本指标体系的科学性,以本研究构建的科学数据质量评价指标体系为基础,使用专家调查法,设计专家调查问卷,通过分析对各级指标计算权重,对评价模型进行调整优化,为我国和地方科学数据中心建立面向决策的科学数据中心评价系统提供参考。

chinaXiv:202004.00024v1

参考文献:

[1] NOAA. NOAA information quality guidelines[EB/OL]. [2019－10－27]. https://www.cio.noaa.gov/services_programs/IQ_Guidelines_103014.html.

[2] DANS. Evaluation of DANS EASY repository based on the FAIR Principles[EB/OL]. [2019－10－27]. https://dans.knaw.nl/en/about/organisation-and-policy/policy-and-strategy/Evaluation-ofDANSEASYbasedontheFAIRprinciples.pdf.

[3] WILKINSON M D, DUMONTIER M, AALBERSBERG I J, et al. Comment: the FAIR guiding principles for scientific data management and stewardship[J]. Scientific data,2016,3:1－9.

[4] 胡聪. 我国科学数据汇交管理现状、问题及对策研究[J]. 科技创业月刊,2019,32(7):81－84.

[5] STVILIA B, GASSER L, TWIDALE M B, et al. A framework for information quality assessment[J]. Journal of the American Society for Information Science and Technology, 2007, 58(12): 1720－1733.

[6] BATINI C, CAPPIELLO C, FRANCALANCI C, et al. Methodologies for data quality assessment and improvement[J]. ACM computing surveys,2009,41(3):1－52.

[7] ZAVERI A, RULA A, MAURINO A. Quality assessment for linked data: a survey[J]. Semantic Web,2016,7(1):63－93.

[8] KAHN M G, RAEBEL M A, GLANZ J M, et al. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research[J]. Medical care, 2012,50(7):S21－S29.

[9] CHEN H, HAILEY D, WANG N, et al. A review of data quality assessment methods for public health information systems[J]. Informational journal of environmental research and public health,2014, 11(5):5170－5207.

[10] HUANG H, STVILIA B, JOERGENSEN C, et al. Prioritization of data quality dimensions and skills requirements in genome annotation work[J]. Journal of the American Society for Information Science and Technology,2012,63(1):195－207.

[11] 邵艳红. 我国政府开放数据质量评价指标体系构建研究[D]. 保定:河北大学,2019.

[12] 李晓彤,翟军,郑贵福. 我国地方政府开放数据的数据质量评价研究——以北京、广州和哈尔滨为例[J]. 情报杂志,2018,37(6):141－145.

[13] 刘桂锋,张裕,刘琼. 科研数据开放平台评价指标体系构建及案例研究[J]. 图书情报知识,2019(1):21－31.

[14] 周宇,廖思琴,阮莉萍,等. 数据监护平台评价指标体系构建与测定研究[J]. 图书馆学研究,2017(1):35－42.

[15] 余芳东. 非传统数据质量评估的国际经验及借鉴[J]. 统计研究,2017,34(12):15－23.

[16] 余厚强,曹雪婷. 替代计量数据质量评估体系构建研究[J]. 图书情报知识,2019(2):19－27,50.

[17] 马费成,望俊成. 信息生命周期研究述评(Ⅰ)[J]. 情报学报,2010(5):939－947.

[18] 丁宁,马浩琴. 国外高校科学数据生命周期管理模型比较研究及借鉴[J]. 图书情报工作,2013,57(6):18－22.

[19] DATAONE. Data life cycle[EB/OL]. [2020－02－09]. https://www.dataone.org/data-life-cycle.

[20] 张洋,肖燕珠. 生命周期视角下《科学数据管理办法》解读及其启示[J]. 图书馆学研究,2019(15):37－43,13.

[21] 国务院办公厅. 国务院办公厅关于印发科学数据管理办法的通知[EB/OL]. [2019－10－26]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.

[22] ORNL DAAC. Data management[EB/OL]. [2019－10－25]. https://daac.ornl.gov/datamanagement/.

[23] UCLA LIBRARY. Documentation and metadata overview[EB/OL]. [2019－10－26]. http://guides.library.ucla.edu/c.php?g=180580&p=1186345.

[24] ARM. Data management plan requirements [EB/OL]. [2019-10-26]. https://www.arm.gov/policies/datapolicies/digitalstatement.

[25] EROS. Data management plans [EB/OL]. [2019-10-25]. https://prd-wret.s3-us-west-2.amazonaws.com/assets/palladium/production/s3fs-public/atoms/files/DMStrategyTemplateVersion1.docx.

[26] USGS. Data management: stewardship [EB/OL]. [2019-10-26] https://www.usgs.gov/products/data-and-tools/data-management/stewardship.

[27] UCLA LIBRARY. About the DMP tool [EB/OL]. [2019-10-26]. http://guides.library.ucla.edu/c.php? g=180580&p=1190077.

[28] NEWCASTLE UNIVERSITY. Research data management: planning [EB/OL]. [2019-10-26]. https://research.ncl.ac.uk/rdm/planning/dmp/writingadatamanagementplan/.

[29] 王丹丹. 科学数据管理计划评价量表分析[J]. 图书情报工作, 2017, 61(18): 35-41.

[30] DEEP BLUE DATA. Policy and terms of use [EB/OL]. [2019-10-26]. https://deepblue.lib.umich.edu/data/agreement#preservation_policy.

[31] UNIPROT. Guidelines for submitting updates or corrections to UniProtKB data [EB/OL]. [2019-10-26]. https://www.uniprot.org/help/submissions.

[32] DIGITAL CONSERVANCY. Policies and guidelines [EB/OL]. [2019-10-26]. https://conservancy.umn.edu/pages/policies/#preservation.

[33] TEXAS DATA REPOSITORY. Digital preservation policy [EB/OL]. [2019-10-25]. https://texasdigitallibrary.atlassian.net/wiki/spaces/TDRUD/pages/291635428/Digital + Preservation + Policy.

[34] 国家基因组科学数据中心. Big standards for big omics data [EB/OL]. [2019-10-26]. https://bigd.big.ac.cn/standards.

[35] 国家基因库生命大数据平台. 提交数据 [EB/OL]. [2019-10-26]. https://db.cngb.org/cnsa/faq/#.

[36] 国家基因库生命大数据平台. 审核数据 [EB/OL]. [2019-10-26]. https://db.cngb.org/cnsa/faq/#.

[37] ORNL DAAC. Submit data [EB/OL]. [2019-10-26] https://daac.ornl.gov/submit/#scope_and_acceptance_policy.

[38] TEXAS DATA REPOSITORY. Metadata dictionary [EB/OL]. [2019-10-25]. https://texasdigitallibrary.atlassian.net/wiki/spaces/TDRUD/pages/493551668/Metadata + Dictionary.

[39] UK DATA SERVICE. Quality assurance [EB/OL]. [2019-10-26]. https://www.ukdataservice.ac.uk/manage-data/format/quality.aspx.

[40] DATAONE. Use data [EB/OL]. [2019-10-26]. https://www.dataone.org/use-data.

[41] DIGITAL CONSERVANCY. About the data repository [EB/OL]. [2019-10-26]. https://conservancy.umn.edu/pages/drum/.

[42] NOAA. NOAA information quality guidelines [EB/OL]. [2019-10-25]. https://www.cio.noaa.gov/services_programs/IQ_Guidelines_103014.html.

[43] USGS. Data management: process and analyze - closely related activities [EB/OL]. [2019-10-25]. https://www.usgs.gov/products/data-and-tools/data-management/process-and-analyze-closely-related-activities? qt-science_support_page_related_con=0#qt-science_support_page_related_con.

[44] TEXAS DATA REPOSITORY. Accessing and evaluating data [EB/OL]. [2019-10-25]. https://texasdigitallibrary.atlassian.net/wiki/spaces/TDRUD/pages/289243266/Accessing + and + Evaluating + Data.

[45] ARM. Data documentation [EB/OL]. [2019-10-26]. https://www.arm.gov/policies/datapolicies/data-documentation.

[46] USGS. Data management: backup & secure [EB/OL]. [2019-10-26]. https://www.usgs.gov/products/data-and-tools/data-management/backup-secure#tools.

[47] TEXAS DATA REPOSITORY. Information security [EB/OL]. [2019-10-25]. https://texasdigitallibrary.atlassian.net/wiki/spaces/TDRUD/pages/292159828/Information + Security.

[48] NEWCASTLE UNIVERSITY. Research data management: working [EB/OL]. [2019-10-26]. https://research.ncl.ac.uk/rdm/working/.

[49] USGS. Data management: archiving [EB/OL]. [2019-10-26]. https://www.usgs.gov/products/data-and-tools/data-management/archiving? qt-science_support_page_related_con=0#qt-science_support_page_related_con.

[50] 黄如花, 邱春艳. 国外科学数据共享研究综述[J]. 情报资料工作, 2013(4): 24-30.

[51] 邢文明, 洪程. 开放为常态, 不开放为例外——解读《科学数据管理办法》中的科学数据共享与利用[J]. 图书馆论坛, 2019(1): 1-8.

[52] USGS. Data management: data release [EB/OL]. [2019-10-26]. https://www.usgs.gov/products/data-and-tools/data-management/data-release#elements.

[53] ICSU WORLD DATASYSTEM. Data sharing principles [EB/OL]. [2019-10-26]. http://www.icsu-wds.org/services/data-sharing-principles.

[54] TEXAS DATA REPOSITORY. Terms of use [EB/OL]. [2019-10-25]. https://texasdigitallibrary.atlassian.net/wiki/spaces/TDRUD/pages/289079299/Terms + of + Use.

作者贡献说明:

江洪: 确定论文选题, 提出修改建议;

王春晓: 论文框架构思, 论文撰写与修改。

chinaXiv:202304.00024v1

ChinaXiv合作期刊

江洪，王春晓.基于科学数据生命周期管理阶段的科学数据质量评价体系构建研究[J].图书情报工作,2020,64(10):19-27.

Research on the Construction of Scientific Data Quality Evaluation System Based on Scientific Data Lifecycle Management Phases

Jiang Hong[1] Wang Chunxiao[1,2]

[1] Wuhan Library, Chinese Academy of Sciences, Wuhan 430071

[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] The evaluation indexes of scientific data quality from 15 scientific data centers at home and abroad are mainly selected in order to screen the common indexes that can objectively reflect the quality of scientific data and build a universal evaluation index model of scientific data quality. [**Method/process**] By using the methods of document investigation, web survey and content analysis, the evaluation indexes of scientific data of 15 scientific data centers were sorted out, and the evaluation indexes of existing scientific data institutions were understood. [**Result/conclusion**] It structures a scientific data quality evaluation index framework based on 5 phases of data lifecycle management, which comprise data management plan, data collection management, data analysis and processing management, data storage management, data sharing and utilization management, and then provides a reference for the establishment of decision-oriented evaluation system of scientific data center in China and local scientific data centers.

Keywords: scientific data data lifecycle evaluation index system index

chinaXiv:202304.00241v1

《图书情报工作》投稿作者学术诚信声明

《图书情报工作》一直秉持发表优秀学术论文成果、促进业界学术交流的使命，并致力于净化学术出版环境，创建良好学术生态。2013 年牵头制订、发布并开始执行《图书馆学期刊关于恪守学术道德净化学术环境的联合声明》(简称《声明》)(见:http://www.lis.ac.cn/CN/column/item202.shtml)，随后又牵头制订并发布《中国图书馆学情报学期刊抵制学术不端联合行动计划》(简称《联合行动计划》)(见: http://www.lis.ac.cn/CN/column/item247.shtml)。为贯彻和落实这一理念，本刊郑重声明，即日起，所有投稿作者须承诺:投稿本刊的论文，须遵守以上《声明》及《联合行动计划》，自觉坚守学术道德，坚决抵制学术不端。《图书情报工作》对一切涉嫌抄袭、剽窃等各种学术不端行为的论文实行零容忍，并采取相应的惩戒手段。

《图书情报工作》杂志社